RESEARCH NOTE

A dynamic fuzzy clustering method based on genetic algorithm*

ZHENG Yan[1**], ZHOU Chunguang[2], LIANG Yanchun[2] and GUO Dongwei[2]

(1. College of Computer Science and Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China; 2. College of Computer Science and Technology, Jilin University, Changchun 130012, China)

Received March 4, 2003; revised May 13, 2003

Abstract A dynamic fuzzy clustering method is presented based on the genetic algorithm. By calculating the fuzzy dissimilarity between samples the essential associations among samples are modeled factually. The fuzzy dissimilarity between two samples is mapped into their Euclidean distance, that is, the high dimensional samples are mapped into the two-dimensional plane. The mapping is optimized globally by the genetic algorithm, which adjusts the coordinates of each sample, and thus the Euclidean distance to approximate to the fuzzy dissimilarity between samples gradually. A key advantage of the proposed method is that the clustering is independent of the space distribution of input samples, which improves the flexibility and visualization. This method possesses characteristics of a faster convergence rate and more exact clustering than some typical clustering algorithms. Simulated experiments show the feasibility and availability of the proposed method.

Keywords: dynamic fuzzy clustering, fuzzy dissimilarity matrix, genetic algorithm, fuzzy c-means clustering.

"Birds of a feather flock together" reveals the meaning of clustering profoundly. Clustering analysis is to classify things in terms of some essential attributes, such that similarity between samples from the same class is significant while similarity between samples from different classes is small. It is a non-supervised pattern recognition problem. The typical clustering algorithms such as fuzzy c-means (FCM)[1], c-means clustering algorithm and so on use samples directly with no preprocessing. Clustering validation depends on the space distribution of the sample considerably[2]. For example, c-means clustering algorithm is suitable only for a hyper-spherical feature space of the sample[3], but not for randomly distributed ones.

We present here a dynamic fuzzy clustering method based on the genetic algorithm (GA), which realizes the clustering for a random feature space of samples. A key advantage of the proposed method is that the clustering is independent of the space distribution of input samples, which improves the flexibility and visualization. More specifically, by constructing a fuzzy dissimilarity matrix, the essential associations among samples is modeled directly such that the high-dimensional samples are mapped into the two-dimensional plane. Then the GA optimizes the coordinates of the samples distributed randomly on a plane, and thus the Euclidean distance between samples approximates to their fuzzy dissimilarity gradually. The GA is a global search algorithm[4], while the FCM and c-means clustering are local search algorithms which tend to converge to a local optimum and fall into the local minimum[5], especially in the case of numerous samples. The proposed dynamic fuzzy clustering method based on GA overcomes the drawbacks mentioned above, improves performance, and shows a faster convergence rate and more exact clustering compared with typical clustering algorithms. Simulated experiments illustrate and examine the feasibility and availability of the proposed method.

1 A dynamic fuzzy clustering method based on GA

1.1 Fuzzy dissimilarity matrix

The fuzzy dissimilarity matrix stores the dissimilarity measurement among the samples. There exist many approaches to constructing the fuzzy dissimilarity matrix, including the quantity product method,

* Supported by the National Natural Science Foundation of China (Grant No. 60175024), the Key Project of Chinese Ministry of Education (No. 02090) and the Key Laboratory for Symbol Computation and Knowledge Engineering of Chinese Ministry of Education

** To whom correspondence should be addressed. E-mail: yanzheng@bupt.edu.cn

the cosine method, the max-min method and the arithmetic average method etc[1].

For the purpose of constructing the fuzzy dissimilarity matrix, the samples must be normalized in the range of [0, 1] in advance. Assume that sample space is $X=\{x_1, x_2, \cdots, x_n\}$. $\forall x_i \in X$ the feature vector is $\boldsymbol{x}_i=(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{ip})$ where x_{ik} denotes the k-th attribute of the i-th sample.

The average and the mean square deviation of the k-th attribute for n samples are, respectively

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}, \tag{1}$$

$$\sigma_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik}-\mu_k)^2}. \tag{2}$$

The initial samples are normalized as follows:

$$x'_{ik} = (x_{ik}-\mu_k)/\sigma_k. \tag{3}$$

Performing the compression using the extremum normalization equation, we have

$$\bar{x}_{ik} = \frac{x'_{ik}-x'_{\min k}}{x'_{\max k}-x'_{\min k}}, \tag{4}$$

where $x'_{\max k}$ and $x'_{\min k}$ are the maximum and the minimum of $x'_{1k}, x'_{2k}, \cdots, x'_{nk}$ respectively.

The fuzzy dissimilarity matrix $(r_{ij})_{mn}$ is a $n\times n$ symmetrical matrix with diagonal elements 1.

$$\begin{bmatrix} 1 & & & & \\ \boldsymbol{r}_{21} & 1 & & & \\ \boldsymbol{r}_{31} & \boldsymbol{r}_{32} & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{r}_{n1} & \boldsymbol{r}_{n2} & \cdots & \cdots & 1 \end{bmatrix}, \tag{5}$$

where r_{ij} denotes the dissimilarity measurement between the samples x_i and x_j, and normally a nonnegative. The closer or more similar x_i and x_j are, the greater the value r_{ij} is; otherwise, the smaller it is.

Using the cosine method the r_{ij} can be written as

$$r_{ij} = \frac{\sum_{k=1}^{p} x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2}\ \sqrt{\sum_{k=1}^{p} x_{jk}^2}}. \tag{6}$$

1.2 Genetic algorithm

After constructing the fuzzy dissimilarity matrix the high-dimensional samples are mapped into the two-dimensional plane. The coordinates of the samples are optimized using the GA, such that the Euclidean distance between samples approximates to their fuzzy dissimilarity. So the error function is defined as

$$E = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=i}^{n} |r'_{ij} - r_{ij}|, \tag{7}$$

where r'_{ij} is the Euclidean distance between the samples x_i and x_j, whose coordinates are (a_i, b_i), $i=(1, 2, \cdots, n)$ and (a_j, b_j), $j=(1, 2, \cdots, n)$ respectively, and r'_{ij} is defined as

$$r'_{ij} = \sqrt{|a_i-a_j|^2+|b_i-b_j|^2}. \tag{8}$$

The smaller the value of the error function is, the greater the fitness of the individual is, and thus the fitness function is defined as

$$f = \frac{1}{E+1}. \tag{9}$$

1.3 Dynamic fuzzy clustering method based on GA

The dynamic fuzzy clustering method based on GA is as follows:

(i) Initialization. Distribute the samples in a plane randomly, i. e. assign randomly coordinate pairs (a_i, b_i) to each sample where $a_i, b_i \in [0, 1]$ $i=1, 2, \cdots, n$.

(ii) Construct the fuzzy dissimilarity matrix $(\boldsymbol{r}_{ij})_{nn}$ using Eqs. (1) ~ (6).

(iii) Form the initial population. Each pair of the coordinates (a_i, b_i) is viewed as a gene and coded to an 8-bit binary. The number of the genes is n. Then the coordinates of all samples are linked into a chromosome (also named as an individual). The length L of a chromosome is $8n$ bits. According to the different order, N individuals are created to form an initial population S.

(iv) Compute the fitness of each individual in S. Calculate the error value of an individual by Eq. (7), and obtain its fitness using Eq. (9).

(v) Select parental individuals. The roulette wheel selection and the elitist strategy are adopted. Firstly, the individual with the greatest fitness is chosen as a parental one. Calculate the selection probabil-

1) Wu, B. Research on swarm intelligence and its application in knowledge discovery. Ph. D. Dissertation, Institute of Computing Technology, Chinese Academy of Sciences, 2002.

ity of each residual individual $p_k = f_k \Big/ \sum_{m=1}^{N} f_m$ and the accumulative probability $q_i = \sum_{j=1}^{i} p_j$. Generate a random even number r in the interval of [0, 1). If $r < q_1$, then select the first individual. Otherwise, if k satisfies the condition of $q_{k-1} \leqslant r < q_k$, then select the individual k. Whirl the roulette wheel for $M-1$ times; thus M individuals might be chosen to form a sub-population S', $S' \subset S$.

(vi) Mate the individuals randomly in S'.

(vii) Crossover operation. Create a random number r in the interval of [0, 1] for each pair of individuals in S'. If $r < p_c$, where p_c is a given crossover probability, then carry out the crossover operation. Next, generate a random number in the interval of $[1, 8n]$ in order to determine the location of crossover, and thus form a sub-population S'' consisting of new individuals.

(viii) Mutation operation. Create a random number r in the interval of [0, 1] for each position of each individual in S''. If $r < p_m$, where p_m is a given mutation probability, then carry out mutation operation in this position.

(ix) Calculate the fitness of all individuals in $S + S''$, and eliminate M individuals with smaller fitness to form a new population S.

(x) Termination. If the fitness of an individual in S satisfies the termination condition, e.g. smaller than ε, then terminate the iterations, decode and obtain the optimal solution; otherwise, return to step (v).

The GA is a subsequent global optimization algorithm, in which genetic operators are used to realize the information transfer, and converges to the global optimal solution. The employment of the elitist strategy and the preprocessing of decreasing the number of the sample dimensions contribute to reduce the required iteration times greatly compared with the FCM.

2 Numerical simulation

For the purpose of illustrating we simulate the proposed algorithm with the test data set of wines from the UCI Machine Learning Database[1] shown in Table 1.

Table 1. The sample data of wines

No.	Attribute												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
2	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
3	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
4	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
...	...	...	...	...	...	...	...	...	...	...	...	...	...
177	13.17	2.59	2.37	20.0	120	1.65	0.68	0.53	1.46	9.30	0.60	1.62	840
178	14.13	4.10	2.74	24.5	96	2.05	0.76	0.56	1.35	9.20	0.61	1.60	560

The test data set consists of 178 instances, 13 attributes and 3 classes. Classes 1, 2 and 3 include 59, 71 and 48 instances respectively. The initialized random distribution of the samples on a plane is shown in Fig. 1.

In the simulations the population size N, the mutation probability p_m and the crossover probability p_c are taken as 150, 0.5 and 0.2 respectively. After 80 iterations the clustering results are obtained and shown in Fig. 2.

Compared with the FCM algorithm the presented method is superior in both recognition accuracy and

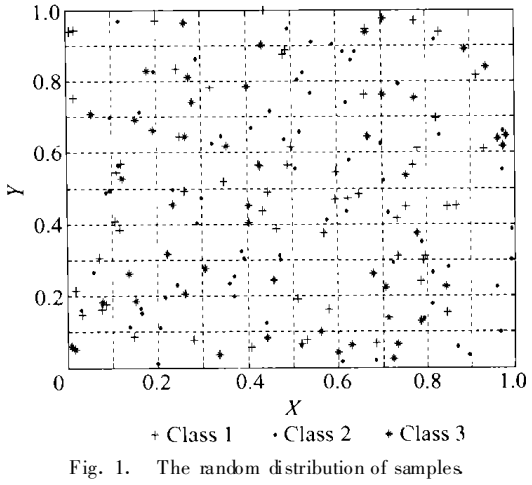

Fig. 1. The random distribution of samples.

1) http://www.ics.uci.edu/~mlearn/MLRepository.html

convergence rate. The comparison of the two methods is shown in Table 2.

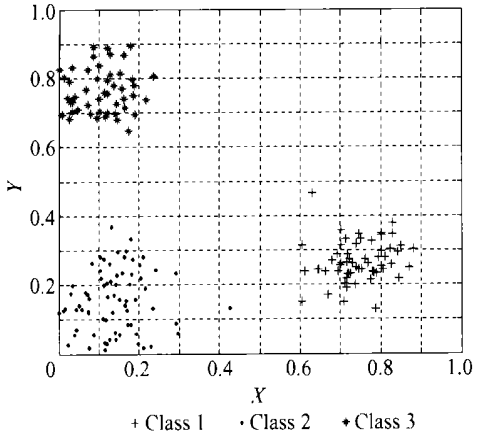

Fig. 2. The clustering results.

Table 2. The comparison of two methods

Algorithm	Iterative times	Accuracy(%)		
		Class 1	Class 2	Class 3
FCM	150	90	92	100
The proposed method	80	93	98	100

3 Conclusions

The validation of existing clustering algorithms depends considerably on the space distribution of the input samples. If the boundary of the sample distribution is evident then the clustering result is satisfactory. However, in real applications the samples may have any distribution forms, which would result in a bad clustering by using the existing methods. In order to solve this kind of problem we present a dynamic fuzzy clustering method based on the GA. With the fuzzy dissimilarity matrix and the GA the high dimensional samples are mapped into the two-dimensional plane, the coordinates of the samples are recursively optimized, and thus the Euclidean distance between the samples approximates to their fuzzy dissimilarity. Simulations show a faster convergence rate and more exact clustering compared with the typical clustering algorithms.

References

1 Huang, F. G. et al. Pattern Recognition (in Chinese). Harbin: Harbin Engineering University Press, 1998.
2 Zhang, L. et al. Kernel clustering algorithm. Chinese Journal of Computers (in Chinese), 2002, 25(6): 587.
3 Gao, X. B. et al. Research development of fuzzy clustering theory and applications. Chinese Science Bulletin (in Chinese), 1999, 21: 2241.
4 Holland, J. H. Adaptation in natural and artificial system. Ann Arbor: University of Michigan Press, 1975.
5 Kamel, S. M. New algorithms for solving the fuzzy c-means clustering problem. Pattern Recognition, 1994, 27: 421.